Elaboração de tesouros utilizando-se o programa de elaboração de tesouros em microcomputador (Tecer)

Lígia Maria Café de Miranda
Marisa Bräscher Basílio Medeiros
Mauro Kenji Sujii

Resumo – Descrevem-se as etapas de elaboração de tesauros utilizando-se o Programa de Elaboração de Tesauros em Microcomputador (TECER). Abordam-se alguns aspectos teóricos sobre tesauros. Apresentam-se as características, os programas e as saídas do Sistema e sugere-se uma metodologia para sua utilização.

1 Introdução

A atividade de elaboração de tesauros é essencialmente intelectual. Procura-se, a partir da análise semântica dos termos, definir relações entre eles e agrupá-los de acordo com a estrutura conceitual do campo do conhecimento que está sendo analisado. Todo esse processo envolve uma tarefa complexa de verificação e controle para que se evite inconsistências.

A complexibilidade estrutural de um tesauro com centenas ou milhares de termos é tão grande que, dificilmente, com métodos puramente manuais, poder-se-á garantir o controle de descritores e não-descritores, a integridade de suas relações, ou mesmo a geração de produtos organizados de forma a que sirvam como instrumentos de apoio à indexação e à busca (listas alfabéticas, listas alfabético-estruturadas, índices permutados, listas por categorias etc). Todos estes problemas indicam a necessidade de um "software" que seja capaz de apoiar a tarefa de criação e manutenção de tesauros. A utilização de um software apropriado permite portanto, a simplificação de diversas atividades na elaboração de tesauros, a sua realização em um menor espaço de tempo, além de garantir um maior grau de confiabilidade dos controles necessários.

O objetivo deste trabalho é descrever as etapas do processo de elaboração de tesauros utilizando-se o Programa de Elaboração de Tesauros em

Microcomputador (TECER), desenvolvido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT).

2 Aspectos Teóricos

Os tesauros são instrumentos de controle terminológico utilizados em sistemas de informação para traduzir linguagem dos documentos, dos indexadores e pesquisadores numa linguagem controlada, usada na indexação e recuperação de informações. Compõem-se por um conjunto de termos e das relações que se estabelecem entre eles.

Os termos eleitos para representarem os conceitos, ou seja, os termos preferidos para a indexação,denominam-se descritores. Os termos denominados não-descritores ou termos proibidos, embora possam descrever os mesmos conceitos que os descritores, não poderão ser utilizados na indexação, para que se evite a proliferação de sinônimos e termos equivalentes.

Existem ainda, além dos descritores e não descritores, termos denominados identificadores e modificadores. Os identificadores representam conceitos individuais, p. ex.: nomes de instituições, planos, projetos. Os modificadores são termos que não são utilizados isoladamente e cuja função é esclarecer ou limitar o significado de descritores. Os qualificadores são um tipo de modificador utilizado para diferenciar homônimos.

As relações estabelecidas nos tesauros são geralmente dos seguintes tipos:

a) *Relação de equivalência:* estabelece-se entre termos que representam o mesmo conceito, ou seja, entre termos sinônimos ou equivalentes. Esses termos são incluídos no tesauro sendo que apenas um deles será o descritor, os outros termos serão considerados não-descritores. Essa relação é expressa pelos símbolos USE e UP (usado para).

Ex.: TECNOLOGIA AGRÍCOLA
UP ENGENHARIA AGRÍCOLA

ENGENHARIA AGRÍCOLA
USE TECNOLOGIA AGRÍCOLA

b) *Relação hierárquica:* exprime os graus de superordenação e subordinação entre os conceitos. O termo superordenado (termo genérico - TG) representa o conceito mais abrangente, do qual o termo subordinado (termo específico – TE) é uma parte ou tipo. Representa-se esta relação através dos símbolos TG e TE.

Ex.: SOLOS
TE SOLOS ÁCIDOS

SOLOS ÁCIDOS
TG SOLOS

Quando um termo específico tem uma relação direta com mais de um termo genérico forma-se um relacionamento *poli-hierárquico*. Alguns tesauros restringem a ligação de um TE a apenas um TG, nesse caso, denominam-se *tesauros mono-hierárquicos*.

c) *Relação associativa:* ocorre entre termos que não são equivalentes nem formam uma hierarquia, mas são tão associados mentalmente que se deve tornar esta ligação explícita no tesauro. O símbolo TR representa esta relação.

Ex.: TRADUTORES
TR TRADUÇÃO

TRADUÇÃO
TR TRADUTORES

Alguns tesauros utilizam a *relação alternativa* para remeter de um não descritor (geralmente termos muito abrangentes) para vários descritores, dos quais somente um deverá ser escolhido para a indexação. Utilizam-se os símbolos VEJA... OU e a recíproca VD (visto de).

Ex.: FLUIDOS
VEJA GASES
OU LÍQUIDOS

GASES
VD FLUIDOS

LÍQUIDOS
VD FLUIDOS

Geralmente os conceitos em um tesauro estão sistematizados em categorias. Definem-se as categorias a partir da análise do(s) campo(s) do conhecimento que está(ão) sendo abordado(s) no tesauro, subdividindo-se este(s) campo(s) em áreas de assunto. Dentre as diversas formas existentes para representar esta divisão, optou-se para efeito de utilização do sistema, pelo estabelecimento de categorias e subcategorias que agrupam conceitos similares, posicionando-os em relação a abrangência temática do tesauro.

Através do estabelecimento de relações e da categorização dos termos é possível delimitar de forma bastante precisa o conceito por ele representado. Entretanto, em alguns casos, torna-se necessária a utilização de notas explicativas (NE) para esclarecer o uso de determinado descritor na indexação.

3 Características Gerais do Programa de Elaboração de Tesauros em Microcomputador (Tecer)

O TECER é um sistema autônomo utilizado na elaboração e manutenção de tesauros, podendo ser considerado uma ferramenta automática de apoio ao processo de elaboração desses instrumentos. Esse sistema foi desenvolvido em Turbo-Pascal versão 3.0 e pode ser executado em qualquer computador

compatível com a linha IBM PC/XT-AT, o que torna acessível à grande maioria das instituições brasileiras.

É um software bastante flexível, dotado de características que oferecem ao usuário diversas possibilidades de estruturação de seu tesauro, garantindo, para todos eles, sua integridade estrutural. Dentre as características mencionadas destacam-se:

a) Gerenciamento de tesauros poli ou mono-hierárquicos;
b) Cadastramento de descritores e não-descritores;
c) Equivalência lingüística: permite que um termo cadastrado no tesauro em um idioma possa ter equivalentes em até três outros idiomas;
d) *Categorização de áreas temáticas:* possibilita a estruturação do tesauro em categorias e subcategorias. Cada categoria representa uma área temática que poderá ou não ser dividida em subcategoria;
e) *Cadastramento de relações:* permite a utilização das relações de equivalência, alternativas, hierárquicas e associativas entre os termos do tesauro, realizando as críticas necessárias para evitar o estabelecimento de relações conflitantes; e
f) *Intercâmbio de tesauros:* possibilita a importação e a exportação de um tesauro e de parte dele. Uma das aplicações deste intercâmbio é a extração de parte de um tesauro multidisciplinar para o desenvolvimento de um tesauro especializado.

Além dessas características, o TECER emite diferentes tipos de saídas para utilização pelo elaborador de tesauro e pelo indexador. São eles:

– *listagem alfabética:* lista os termos do tesauro em ordem alfabética;

– *listagem alfabético-estruturada:* lista os termos em ordem alfabética acompanhados de suas relações semânticas e notas explicativas;

– *saída hierárquica:* apresenta as cadeias hierárquicas em ordem alfabética, sendo que o termo de entrada é o descritor mais genérico da cadeia;

– *índice permutado:* índice alfabético no qual cada palavra significativa que compõe o descritor constitui-se numa entrada, localizada numa posição fixa da página, seguida de outras palavras do descritor;

– *relatórios estatísticos:* dados estatísticos sobre: número total de termos, número total de descritores e não-descritores, número de relações hierárquicas, associativas, alternativas e de equivalência, média de relações por descritor e média de subcategorias por categoria. Com estes dados, pode-se estimar o tamanho do tesauro para efeito de intercâmbio e para seu gerenciamento;

– *relatórios das alterações:* relaciona as transações efetuadas durante o processo de elaboração e manutenção do tesauro.

As rotinas do TECER são executadas através dos seguintes programas:

a) *CRIAÇÃO:* utilizado para definir os parâmetros para criação de um tesauro, como: nome do tesauro, idiomas utilizados, se tesauro mono-hierárquico ou poli-hierárquico etc.
b) *TERMOS:* inclui, exclui, altera e lista o nome de termos no(s) idioma(s) do tesauro.
c) *CLASSES:* inclui, exclui, altera e lista as categorias e subcategorias do tesauro. Esse programa executa também o tratamento de termos por categorias e subcategorias.
d) *RELAÇÕES:* inclui, exclui e lista as relações dos termos do tesauro, executando as críticas necessárias.
e) *ALTERADO:* emite relatórios de alteração.
f) *SAÍDAS:* gera os diversos tipos de saídas do sistema.
g) *EXPORTA:* executa o processo de exportação de tesauros.
h) *IMPORTA:* executa o processo de importação de tesauros.
i) *ÚTEIS:* executa rotinas utilitárias do sistema, como: reorganização de arquivos, manipulação do arquivo de palavras vazias, etc.

4 Metodologia

Na elaboração de tesauros utilizando-se o TECER, após o planejamento inicial, onde serão estabelecidas as características gerais do tesauro, deverão ser definidos determinados parâmetros no programa CRIAÇÃO, que implicam em diferentes tipos de tratamento pelo sistema:

a) se o tesauro será poli ou mono-hierárquico. Essa definição é necessária para que sejam efetuadas críticas específicas a cada tipo de tesauro no programa RELAÇÕES;

b) se o tesauro será multilingüe ou monolingüe para que sejam criados arquivos de termos nos diversos idiomas quando se tratar de tesauro multilingüe;

e, no caso de tesauro estruturado em categorias:

c) se será permitida a inclusão de um termo em mais de uma categoria ou se um termo poderá ser cadastrado em apenas uma categoria; e

d) se o tesauro será estruturado tematicamente em categorias e subcategorias ou apenas em grandes categorias.

Estando esses parâmetros definidos, sugere-se, nos itens seguintes, uma seqüência de tarefas a serem realizadas durante o processo de elaboração de tesauros.

4.1 Tratamento de Termos

A partir da coleta de termos realizada pela equipe de elaboração do tesauro, independentemente da metodologia adotada, deverá ser realizada a inclusão dos termos em máquina utilizando-se o programa TERMOS, de acordo com as seguintes etapas:

a) Incluir os termos, aos quais serão atribuídos códigos de identificação automaticamente pelo sistema. No caso de tesauros multilingües poderão ser cadastrados os termos em todos os idiomas do tesauro simultaneamente ou, após a inclusão do termo do idioma de trabalho, incluir-se os equivalentes à medida que forem sendo identificados;

b) Emitir saída alfabética dos termos incluídos para verificação (Programa SAÍDAS);

c) Analisar as listagens para seleção dos termos que permanecerão inicialmente no tesauro, para padronização dos termos de acordo com as regras estabelecidas para entrada de descritores e para identificação de termos ausentes;

d) Excluir os termos considerados irrelevantes para o tesauro, alterar os termos para a forma padronizada e incluir novos termos (Programa TERMOS);

e) Emitir nova listagem alfabética para conferir as alterações efetuadas (Programa SAÍDAS); e

f) Emitir relatório de alterações para registro das operações realizadas (Programa SAÍDAS).

As etapas de inclusão, exclusão e alteração de termos, bem como, a geração de listagens para correções serão realizadas até que se chegue à listagem dos termos que comporão o tesauro.

Caso o tesauro não seja estruturado em categorias, deve-se, após concluída esta etapa, proceder ao cadastramento das relações entre os termos (item 4.3).

4.2 Tratamento de Categorias

Para o tratamento temático dos termos podem ser cadastradas categorias e subcategorias. Caso tenha se optado pela estrutura temática apenas por categorias, os termos serão agrupados dentro das categorias cadastradas.

Optando-se pela utilização de subcategorias os termos serão agrupados dentro delas e as categorias servirão apenas para agrupar as subcategorias pertencentes à ela. Só é possível cadastrar um nível de subcategorias.

Em tesauros que utilizam identificadores, estes podem compor uma categoria e, caso necessário, serem divididos em subcategorias.

Após o cadastramento das categorias e, quando for o caso, das subcategorias deve-se efetuar o agrupamento de termos utilizando-se a listagem alfabética. Para inclusão em máquina utiliza-se o Programa CLASSES.

A partir do agrupamento dos termos torna-se mais fácil a identificação das relações conceituais que se estabelecem entre eles, principalmente as relações hierárquicas e de equivalência, pois os termos que se relacionam estarão próximos de acordo com a estrutura temática já definida.

4.3 Tratamento de Relações

Para o tratamento semântico dos termos podem ser estabelecidos os seguintes tipos de relações: relações de equivalência, relação alternativa, relação hierárquica e relação associativa.

As notas explicativas são incluídas no sistema como um tipo de relação. São cadastradas à medida em que se sente a necessidade de esclarecer o uso de um descritor.

Para o estabelecimento das relações utiliza-se o programa RELAÇÕES. Sugere-se que sejam identificadas e cadastradas em máquina inicialmente as relações de equivalência e alternativas. Só é necessário cadastrar a relação uma vez, a recíproca será gerada automaticamente. Dessa forma o programa já reconhecerá quais serão os descritores e não descritores do tesauro, e não permitirá o estabelecimento de relações hierárquicas e associativas com não-descritores.

Após o cadastramento das relações de equivalência e alternativas emite-se uma listagem alfabético-estruturada (programa SAÍDAS) e, a partir dessa listagem, estabelecem-se as relações hierárquicas.

Sugere-se que cada termo seja analisado, identificando-se qual o seu termo genérico (TG). Os termos que não possuem TG serão os termos do topo da cadeia hierárquica.

Finalizando-se a inclusão de relações hierárquicas, emite-se uma listagem alfabético-estruturada na qual já será possível verificar-se as relações de equivalência, alternativas e hierárquicas, bem como, as notas explicativas já cadastradas. Analisando-se esta listagem definem-se as relações associativas para posterior inclusão em máquina.

Incluindo-se todas as relações, conclui-se a fase de entrada dos dados no sistema, restando apenas efetuar alterações onde for necessário. Para análise do tesauro no todo podem ser gerados os tipos de saídas, descritos no item 3, dependendo do objetivo da análise.

5 Conclusão

A elaboração de tesauros, apesar de ser uma tarefa de cunho intelectual, pode ser facilitada a partir da utilização de recursos automáticos. O Programa de Elaboração de Tesauros em Microcomputador (TECER) foi desenvolvido com este objetivo. Este Sistema permite a simplificação das diversas atividades de elaboração de tesauros, a sua realização em um menor espaço de tempo, além de garantir um maior grau de confiabilidade dos controles necessários.

Todos os módulos do TECER foram testados em tarefas reais e sintonizadas com necessidades práticas do dia a dia operacional do elaborador de tesauro. Os resultados de desempenho e conforto de operação durante o desenvolvimento do Tesauro de Ciência da Informação (TCI) são a garantia de que o sistema pode ser aplicado com sucesso à elaboração de tesauros inclusive de porte significativamente maior.

6 Referências Bibliográficas

1. IBICT. *Diretrizes para elaboração de tesauros monolingües.* Brasília, 1984. 70p.
2. IBICT. [*Manual de utilização:* software tesauro.] Brasília, 1988.